

Partition of Random Items: Tradeoff between Binning Utility and Meta Information Leakage

Farhang Bayat, Shuangqing Wei

Abstract—In this paper, we propose a novel formulation to understand the tradeoff between binning utility and meta information leakage when we face the problems of partitioning random items. As an example, such problems could emerge when online users attempt to protect their browsing behavior patterns to certain extent by resorting to multiple proxy websites. Under the framework, we formulate a constrained optimization problem where the goal is to maximize binning utility while restraining a certain level of information leakage by properly dividing a set of M random items into $N = 2$ bins. We then offer algorithms and complexity orders to solve such a problem. These developments will be the cornerstone for our further investigation of more generalized scenarios.

Index Terms—partition, information leakage, privacy, mutual information, submodularity

I. INTRODUCTION

Today internet has become so intertwined with our everyday activities that it is impossible to imagine living without it. However, this dependency could be subject to exploitation by the eavesdroppers. Assuming every browse as a query, it could be argued that the search history of a user contains a series of queries specific to him which could point to his specific likes and dislikes. By following every user's history of browses vital information specific to each user could be developed. It then becomes important that each user attempt to add some level of protection to their browses to hide necessary -or rather enough information about themselves.

To do so, the user has some options. First option is for him to use proxy softwares also known as VPNs; such programs fully change the user's IP address rendering him undetectable and thus futile for the service provider to track. The problem with such softwares is the fact that they slow down the connection -such as the incognito mode in different browsers. Thus, while all the information does stay hidden, a reduction of speed prevents this method from being highly desirable.

Another method is the use of proxy websites. Such websites offer the user a URL box where he can input any website he wishes to visit. The only difference is that in such websites, the address input is decoded into a series of characters which appear at the end of the URL of the original proxy website. These characters change through time by seconds meaning if the service provider records the URL opened through the proxy website and decides to open it to access specific content by inputting the URL, he will

not go to the decoded web page. Still, such websites slow the connection. However, through this method, the user has the option of choosing multiple proxy websites and thus presumably cutting down on the utility loss. Thus if a utility function based upon connection speed -bandwidth- for the user is calculable, a privacy constrained problem between the user and an eavesdropper (for example a service provider) could be defined. The solution to such a problem could offer insights in regards to the tradeoff between proxy allocation utility and meta information leakage when we face the problems of partitioning a set of random items (i.e. websites that a user has chosen to visit following his own distributions) into a given number of bins (i.e. a given set of proxy servers each of which has its own utility function, as will be further detailed in Section II).

In our proposed framework as detailed in Section II, meta data information refers to the patterns about a sequence of items (for example users favorable websites) infer-able based upon a sequence of bins (e.g. proxy sites) observed by an eavesdropper. This assumption is an expansion of [1] and [2]. However, it should be noted that due to usage of proxy sites, an eavesdropper cannot directly observe the original and input items, but rather bin indexes. Under our proposed novel framework, we further propose two approaches, based upon exhaustive search and submodular functions, respectively, to solving the formulated constraint optimization problem, namely, dividing M random items into $N = 2$ bins, under an upper-bound on leaked meta information.

This paper represents an expansion of our previous works in [3], [4] where in [3] a utility function was introduced in the form of the average stopping time for detection of an active subgraph using certain queries while in [4], we developed a concept of information leakage through a vast set of possible queries. In this paper, we shift our attention from detection of an active subgraph to seeking tradeoff between optimizing utility of partitioning a set of random items and restraining information leakage. From the perspective of [3,4], a set of random items could represent a set of edges in a given subgraph, which become active probabilistic-ally, while a partition strategy corresponds to a particular option of querying edges.

The rest of the paper is organized as follows. In Section II we formulate the problem in terms of privacy and utility functions. We dissect what the goal and the constraints are. We then explain why due to imposed complexities we have chosen to simplify the problem to only 2 bins in

⁰F. Bayat and S. Wei are with the School of EECS, Louisiana State University, Baton Rouge, LA 70803, USA (Email: swei@lsu.edu; fbayat1@lsu.edu). This work has been supported in part by the National Science Foundation under Grant No. 1320351.

Section ???. In Section ??, we offer a full algorithm to an exhaustive search solution to a specific expansion of our problem; multiple algorithms are further developed and their corresponding complexities are addressed. In Section III, we show the conditions under which both the objective and constraint functions are equipped with submodular property under which we can leverage efficient algorithms to seek the optimal partition of random items for more general scenarios. Finally, conclusions are drawn in Section ??.

II. SYSTEM MODEL

First, we propose an abstract framework to formalize the goal of seeking partition of N items into M bins. More specifically, we aim to allocate each one of $1 \leq i \leq M$ possible items (queries) to one of N output bins. There could be at most N^M such partitions. It follows that any set allocation $A_l, 1 \leq l \leq N^M$ results in N sets $S_j^{(l)} \subseteq \{1, 2, \dots, M\}, j = 1, 2, \dots, N$. Each such set is defined as

$$S_j^{(l)} = \{i | \theta_{i,j} = 1\}$$

$$\text{where } \theta_{i,j} = \begin{cases} 1 & i \in S_j^{(l)} \\ 0 & i \notin S_j^{(l)} \end{cases} \quad (1)$$

We further assume $S_j^{(l)} \cap S_k^{(l)} = \emptyset, j \neq k$. Furthermore we have $\bigcup_{j=1}^N S_j^{(l)} = \{1, 2, \dots, M\}$. Finally the size of each such set $S_j^{(l)}$ is defined as $L_j^{(l)}$.

A. Probabilistic Model

We assume at any time slot one and only one of the inputs is chosen with a certain probability. Thus if we use variable $X \in \{1, 2, \dots, M\}$ as a representation of set of items, we could have $P(X = i) = P(\gamma_i = 1) = \pi_i, 1 \leq i \leq M$ as a representation of the probability of choosing item i from the set X where $\gamma_i \in \{0, 1\}$. It further follows that $\sum_{i=1}^M \gamma_i = 1$, stipulating that one and only one of M items is selected. These M items could represent a set of M web pages to be visited by a user at a particular time instant. The prior probability distribution of M items reflects the user's favoritism toward these web pages.

Next, we introduce an observable random variable $Y \in \{1, 2, \dots, N\}$, denoting the index of the bin (the proxy site) employed to carry one of the $M > N$ items. It follows that the probability of each bin's appearance given a set allocation scheme such as A_l will be equal to $P(Y = j | A_l) = \sum_{i=1}^M P(Y = j | A_l, X = i) P(X = i | A_l) = \sum_{i=1}^M P(Y = j | A_l, X = i) P(X = i)$ where we have dropped the second conditional probability due to the independence between X and A_l . Furthermore $P(Y = j | A_l, X = i) = \theta_{i,j}^{(l)} \in \{0, 1\}$. It thus follows that

$$P(Y = j | A_l) = \sum_{i=1}^M \theta_{i,j}^{(l)} P(X = i) \rightarrow$$

$$P(Y = j | A_l) = \sum_{i \in S_j^{(l)}} \pi_i = \alpha_j^{(l)} \quad (2)$$

B. Revealed Information

By choosing to allocate M items to N bins where $N \leq M$, we have injected ambiguity and uncertainty into the output binning index sequence about the input item sequence over a successive n visits or channel uses. In other words, if we originally chose to transmit n of such items, our total set of possible sequences would be of form $\vec{X}^n = [X_1 X_2 \dots X_n]$ out of M^n possible outcomes. From an observer's perspective which can only have access to which one of N bins is deployed in each time slot, sequences in the form of $\vec{Y}^n = [Y_1 Y_2 \dots Y_n]$ has cardinality of at most $N^n < M^n$. Despite the amount of uncertainty added due to the many-to-one mapping between items and bins, the output sequence still reveals certain amount of information regarding the patterns of sequences of M random items.

This observation could be further studied by indicating how our allocation system resembles a coding framework where we have an equivalent channel whose input variable is X and output Y , as show in Figure 1.

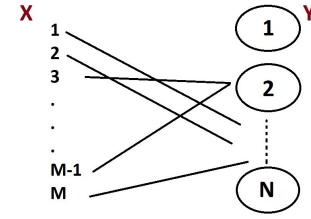


Figure 1. Coding Channel Representation of the Problem

Under such a framework, the equivalent channel output sequence $\vec{Y}^n$ can help an eavesdropper classify the input sequence $\vec{X}^n$ into a number of differential classes. As a result, information about the specific input item patterns is leaked to certain degree and can be measured using conditional mutual information $I(X; Y | A_l)$ between X and Y , given a particular channel mapping (i.e. partition A_l relationship as illustrated in Figure 1).

Such conditional mutual information thus measures the maximum number of bits of meta information about item sequence per channel use. Therefore, we can have at most $2^{nI(X; Y | A_l)}$ sequences $\vec{X}^n$ distinguishable by inferring based on $\vec{Y}^n$. We thus adopt $I(X; Y | A_l)$ as the privacy metric conditioned on a particular partition mapping A_l .

It follows that due to the combinatorial nature of a set allocation problem there are a total of N^M possible methods to allocate these M items to the N sets. We can formulate the mutual information over a set allocation $A_l, 1 \leq l \leq N^M$ as:

$$I(X; Y | A_l) = H(X | A_l) - H(X | Y, A_l) =$$

$$H(Y | A_l) - H(Y | X, A_l) = H(Y | A_l) =$$

$$H(\alpha_1^{(l)}, \alpha_2^{(l)}, \dots, \alpha_N^{(l)}) \quad (3)$$

where we have used the notion of $H(a_1, a_2, \dots, a_m) = -\sum_{v=1}^m a_v \log a_v$ and the fact that $H(Y | X, A_l) = 0$ because

if both the input X and the channel scheme A_l are known, then output Y could simply be calculated.

C. Utility Function

Note that the main reason why we chose bin allocation was to reach a higher utility function. In this section we define a utility function to apply to the problem. If an allocation scheme A_l has resulted in $S_1^{(l)}, S_2^{(l)}, \dots, S_N^{(l)}$, we assume each of the N bins offer a utility function of their own based upon the set they have been bestowed. Every bin j thus offers a utility represented by $f_j(S_j^{(l)})$. It then follows that the overall utility function will be in the form of $U_l = \sum_{j=1}^N f_j(S_j^{(l)})P(Y = j|A_l) = \sum_{j=1}^N \alpha_j^{(l)} f_j(S_j^{(l)})$.

D. Problem Definition

Based on the previous observations we set a goal to find the set allocation A_l over which (a) U_l is maximized and (b) $I(X; Y|A_l) \leq I_{th}$ where I_{th} represents the maximal allowed revealed information.

We aim to find the set allocation scheme A_l which helps

$$\begin{aligned} \max_{1 \leq l \leq N^M} U_l &= \sum_{j=1}^N \alpha_j^{(l)} f_j(S_j^{(l)}) \\ \text{while } H(\alpha_1^{(l)}, \alpha_2^{(l)}, \dots, \alpha_N^{(l)}) &\leq I_{th} \\ \text{and } \alpha_j^{(l)} &= \sum_{i \in S_j^{(l)}} \pi_i \end{aligned} \quad (4)$$

As shown in Eq. 4, we in general have to deal with an exponentially large number N^M of candidates, under a given N and with respect to M .

III. MULTI-SUBMODULAR SET FUNCTIONS AS A MEANS OF SOLUTION

As mentioned previously, the problem formulated in Eq. (4) is NP-complicated. Still, we could opt to utilize the definition of multi-submodular set functions so as to reduce the complexity to that of polynomial at the cost of error. In order to do so, we first introduce a new realization of the problem. We then express concerns about possible solutions. Next, we offer insight as to how we could deal with each case.

A. New Problem Development

We aim to gather both the utility and the constraint imposed in Eq. (4) in the form of one function we hope to maximize. Thus, a new maximization problem could be developed:

$$\max_{1 \leq l \leq N^M} U_l + \lambda(I_{th} - H(\alpha_1^{(l)}, \alpha_2^{(l)}, \dots, \alpha_N^{(l)})) \quad (5)$$

where $\lambda \geq 0$ represents a variable connecting the utility and the constraint to each other so as to allow comparison

between them. We can further express this equation by opening it:

$$\max_{1 \leq l \leq N^M} \sum_{j=1}^N [\alpha_j^{(l)} f_j(S_j^{(l)}) + \lambda \alpha_j^{(l)} \log(\alpha_j^{(l)})] \quad (6)$$

We can now see that if we define $F_j^{(l)} = \alpha_j^{(l)} f_j(S_j^{(l)}) + \lambda \alpha_j^{(l)} \log(\alpha_j^{(l)})$, Equation (6) is simply a sum of functions defined over a series of sets. We refer to these as multi-variate set functions seeing as how their values are based upon specific sets and variables introduced in each of these sets.

B. Imposing Multi-submodularity

[5] first proved that if we can prove multi-submodularity for functions such as those formulated in Eq. (6), then they could be modeled as simpler problems (submodular set functions). We thus, aim to find the sufficient conditions for such occurrence. To do so, we first offer a definition of multi-submodularity.

As mentioned in [5], a multivariate function $F : N^M \rightarrow R$ is multi-submodular if for all pairs of tuples $(S_1, \dots, S_M)$ and $(T_1, \dots, T_M) \in N^M$ we will have:

$$\begin{aligned} F(S_1, \dots, S_M) + F(T_1, \dots, T_M) &\geq F(S_1 \cup T_1, \dots, S_M \cup T_M) \\ &\quad + F(S_1 \cap T_1, \dots, S_M \cap T_M) \end{aligned} \quad (7)$$

Since in our formulation functions are separately defined on different sets, the condition in Eq. (7) is simplified to the sufficient condition of submodularity of $F_j^{(l)}$ for all sets S_j for Equation (6). We now need to find the sufficient condition for submodularity of $F_j^{(l)}$ when defined over a set S_j .

C. Imposing Separate Submodularities

For an easier mathematical representation of the following derivation we assume $F(S_j) = F_j^{(l)}$. Without loss of generality we also assume that $f_j(S_j^{(l)}) = f(S_j^{(l)})$. The first assumption simply dictates that once a mapping scheme A_l is chosen, its index does not play any further role in our proof. The second assumption necessitates the utility function defined over all sets to be the same.

In the next step, we opt to use diminishing return property as the means of making certain each of these functions are submodular. Following is a definition of diminishing returns for submodular functions, after which we derive the sufficient conditions for the case discussed in Eq. (6).

Diminishing Property Return dictates that a set function $F : 2^\Omega \rightarrow R$ is submodular if, for all $A, B \subseteq \Omega$ with $A \subseteq B$ and for each $x \in \Omega - B$ we have [6]:

$$F(A \cup \{x\}) - F(A) \geq F(B \cup \{x\}) - F(B) \quad (8)$$

Now we attempt to expand Eq. (8) for each $F(S_j)$. However, to properly do so, we first need to account for the behavior of this function.

We have defined $F(S_j) = \alpha_j f(S_j) + \lambda \alpha_j \log(\alpha_j)$ where it seems that the function has a singular relationship with set

S_j . However; there is a secondary relationship the function shares with the set $S_j^C = S - S_j$ where S represents the universal set. This relationship could be modeled as $F(S_j^C) = (1 - \beta_j)f(S - S_j^C) + \lambda(1 - \beta_j)\log(1 - \beta_j)$ where we have used the fact that $\beta_j = 1 - \alpha_j$ seeing as how we define

$$\beta_j^{(l)} = \sum_{i \in \{S - S_j^{(l)}\}} \pi_i \quad (9)$$

Thus, for any set S_j , we must find the sufficient conditions for the existence of diminishing property for both functions $F_1(S_j) = \alpha_j f(S_j) + \lambda \alpha_j \log(\alpha_j)$ and $F_2(S_j) = (1 - \alpha_j)f(S - S_j) + \lambda(1 - \alpha_j)\log(1 - \alpha_j)$. To do so, we will calculate their necessary conditions and then find their $\cap$ as the final conditions (assuming they do not negate one another).

Note: For any further references, we first need to address a series of variable and function definitions which are going to play a vital role in the rest of this paper:

Definitions

1. Any variable represented with a capital Letter represents a set.
2. Any variable represented with a small letter represents an element.
3. α_x represents the probability of item x and α_A represents the sum of probabilities of items mapped into a set A .
4. α_{BA} represents the difference in the sum of probabilities of items mapped into the sets B and A which could be further shown as $\alpha_{BA} = \alpha_B - \alpha_A$.
5. $g(C)$ represents the 1st order difference of a set function $f(c)$ which could be formulated as $g(C) = f(C) - f(C - D)$.
6. $q(C)$ represents the 2nd order difference of a set function $f(c)$ which could be formulated as $q(C) = g(C) - g(C - D)$.
7. We assume the probability of items is sorted in a decreasing manner as $\pi_1 \geq \pi_2 \geq \dots \geq \pi_M$.

Lemma III.1. *The function $F_1(S_j)$ is submodular if*

- (1) $g(C) \leq 0$
 - (2) $q(C) \leq 0$
 - (3) $|g(C)| \geq \lambda \log(\frac{1}{\pi_M})$
- for all sets C .

The proof for this lemma is presented in Appendix.

Lemma III.2. *The function $F_2(S_j)$ is submodular if*

- (1) $g(C) \leq 0$
 - (2) $q(C) \leq 0$
 - (3) $|g(C)| \geq \lambda \log(\frac{1}{\pi_M})$
- for all sets C .

The proof for this lemma is presented in Appendix. We can then use the results from two lemmas above to conclude that:

Corollary III.2.1. *The function $F(S_j)$ is submodular if*

- (1) $g(C) \leq 0$

- (2) $q(C) \leq 0$
 - (3) $|g(C)| \geq \lambda \log(\frac{1}{\pi_M})$
- for all sets C .

Unfortunately, [5] does not provide us with an algorithm to remodel our multi-submodular problem in a submodular problem, they simply prove that this could be done. Thus, in order to expand upon the idea of polynomial complexity of solution algorithms we opt to assume $N = 2$ and offer the reader the algorithm to deal with such a specific case. We then calculate the complexity imposed by the algorithm to further stress the benefits of using such an idea in spite of accepting error.

Lemma III.3. *When $N = 2$, the function in Eq. (5) is submodular if*

- (1) $g(C) \leq 0$
 - (2) $2|g(C)| \geq \lambda \log(K), K < (\frac{1}{\pi_M})^2$
- for all sets C .

REFERENCES

- [1] Y. Chen, C. Suh, and A. J. Goldsmith, "Information recovery from pairwise measurements," *IEEE Transactions on Information Theory*, vol. 62, no. 10, pp. 5881–5905, Oct 2016.
- [2] E. Zheleva and L. Getoor, *Preserving the Privacy of Sensitive Relationships in Graph Data*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 153–171. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-78478-4_9
- [3] F. Bayat and S. Wei, "Sequential detection of disjoint subgraphs over boolean mac channels: A probabilistic approach," in *2016 IEEE Globecom Workshops (GC Wkshps)*, Dec 2016, pp. 1–6.
- [4] —, "Non-adaptive sequential detection of active edge-wise disjoint subgraphs under privacy constraints," in *IEEE Transactions on Information Forensics Security*, Submitted October 2017.
- [5] R. Santiago and F. B. Shepherd, "Multi-agent and multivariate submodular optimization," *CoRR*, vol. abs/1612.05222, 2016.
- [6] W. J. Cook, W. H. Cunningham, W. R. Pulleyblank, and A. Schrijver, *Combinatorial Optimization*. New York, NY, USA: John Wiley & Sons, Inc., 1998.

APPENDIX

A. Lemma III.1

Proof. Starting for $F_1(S_j)$, by rewriting Eq. (8) we will have:

$$\begin{aligned} &(\alpha_A + \alpha_x)f(A + \{x\}) + \lambda(\alpha_A + \alpha_x)\log(\alpha_A + \alpha_x) \\ &\quad - \alpha_A f(A) - \lambda \alpha_A \log(\alpha_A) \geq \\ &(\alpha_A + \alpha_{BA} + \alpha_x)f(B + \{x\}) \\ &\quad + \lambda(\alpha_A + \alpha_{BA} + \alpha_x)\log(\alpha_A + \alpha_{BA} + \alpha_x) \\ &\quad - (\alpha_A + \alpha_{BA})f(B) - \lambda(\alpha_A + \alpha_{BA})\log(\alpha_A + \alpha_{BA}) \end{aligned}$$

We can then factorize the inequality as

$$\begin{aligned} &\alpha_{BA}[-f(B + \{x\}) + f(B) - \lambda \log(\alpha_A + \alpha_{BA} + \alpha_x) \\ &\quad + \lambda \log(\alpha_A + \alpha_{BA})] + \alpha_x[f(A + \{x\}) \\ &\quad - f(B + \{x\}) + \lambda \log(\alpha_A + \alpha_x) - \lambda \log(\alpha_A + \alpha_{BA} + \alpha_x)] \\ &\quad + \alpha_A[f(A + \{x\}) - f(B + \{x\}) - f(A) + f(B) \\ &\quad + \lambda \log(\alpha_A + \alpha_x) - \lambda \log(\alpha_A + \alpha_{BA} + \alpha_x) \\ &\quad - \lambda \log(\alpha_A) + \lambda \log(\alpha_A + \alpha_{BA})] \geq 0 \end{aligned}$$

We assume $g(x+y) = f(x+y) - f(x)$, $q(x+y) = g(x+y) - g(x)$, $y \geq 0$. We can then rewrite the factorization as:

$$\begin{aligned} & \alpha_{BA}[-g(B + \{x\}) + \lambda \log \left[\frac{1}{1 + \frac{\alpha_x}{\alpha_A + \alpha_{BA}}} \right]] \\ & + \alpha_x[-g(B + \{x\}) + \lambda \log \left[\frac{1}{1 + \frac{\alpha_{BA}}{\alpha_A + \alpha_x}} \right]] \\ & + \alpha_A[-q(B + \{xx\}) + \lambda \log \left[\frac{1 + \frac{\alpha_A}{\alpha_{BA}}}{1 + \frac{\alpha_{BA}}{\alpha_A + \alpha_x}} \right]] \geq 0 \end{aligned}$$

Since the above inequality needs to hold true for all possible sets of $A \subseteq B$, $x \notin B$, we aim to determine the maximal amount enforced by the above set of inequalities. The first factorization results in two inequalities: (1) $g(C) \leq 0$ for all sets C and (2) $|g(C)| \geq \lambda \max(\log(1 + \frac{\alpha_x}{\alpha_B}))$. To find the maximum of such a limit, we need to impose the one item with highest probability to $\{x\}$ and assume the one lowest probability item to set B . Then the above inequality is maximized.

It is important to note that by doing so, we are removing the possibility of $B = \emptyset$ as a case. In other words, we are assuming that $\alpha_B = 0$ is not a scenario we need to discuss. We will now demonstrate why such an assumption is accurate.

When $B = \emptyset$, we can conclude that $\alpha_A = \alpha_B = 0$ and then rewrite inequality (8) in the following format:

$$\alpha_x f(\{x\}) \geq \alpha_x f(\{x\})$$

which is always true.

The second factorization results in two inequalities: (1) $g(C) \leq 0$ for all sets C and (2) $|g(C)| \geq \lambda \max(\log(1 + \frac{\alpha_{BA}}{\alpha_A + \alpha_x}))$. To find the maximum of such a limit, we need to impose the item with lowest probability to $\{x\}$ and that $\alpha_A = 0$ and then have $\alpha_{BA} = 1 - \alpha_x = \alpha_B$.

Once again, it is important to note that by doing so, we are removing the possibility of $\alpha_A = 0$ and $\alpha_x = 0$ as a case. In other words, we are assuming that $\alpha_A + \alpha_x = 0$ is not a scenario we need to discuss. We will now demonstrate why such an assumption is accurate.

When $\alpha_A + \alpha_x = 0$, we can rewrite inequality (8) in the following format:

$$0 \geq \alpha_{BA} f(B) - \alpha_{BA} f(B) = 0$$

which is always true.

The third factorization could be simplified. The logarithm argument consists of a nominator greater than denominator thus resulting in the overall logarithm argument to be positive. We thus only need to impose that $q(C) \leq 0$. It then follows that the probability of each item is sorted in a decreasing manner as $\pi_1, \dots, \pi_M$, we can write the set of 3 sufficient conditions for submodularity of set function $F_1(S_j)$, $j = 1, \dots, N$ as (1) $g(C) \leq 0$ (2) $q(C) \leq 0$ (3) $|g(C)| \geq \lambda \log \left[\max(1 + \frac{\pi_1}{\pi_M}, 1 + \frac{1 - \pi_M}{\pi_M}) \right] = \lambda \log \left(\frac{1}{\pi_M} \right)$. $\square$

B. Lemma III.2

Proof. We now follow the same method for $F_2(S_j)$, $j = 1, \dots, N$ by rewriting Eq. (8):

$$\begin{aligned} & (1 - \alpha_A - \alpha_x)f(S - A - \{x\}) + \\ & \lambda(1 - \alpha_A - \alpha_x) \log(1 - \alpha_A - \alpha_x) - (1 - \alpha_A)f(S - A) \\ & - \lambda(1 - \alpha_A) \log(1 - \alpha_A) \geq \\ & (1 - \alpha_A - \alpha_{BA} - \alpha_x)f(S - B - \{x\}) \\ & + \lambda(1 - \alpha_A - \alpha_{BA} - \alpha_x) \log(1 - \alpha_A - \alpha_{BA} - \alpha_x) \\ & - (1 - \alpha_A - \alpha_{BA})f(S - B) - \\ & \lambda(1 - \alpha_A - \alpha_{BA}) \log(1 - \alpha_A - \alpha_{BA}) \end{aligned}$$

We can then factorize the inequality as

$$\begin{aligned} & \alpha_{BA}[f(S - B - \{x\}) - f(S - B) \\ & - \lambda \log(1 + \frac{\alpha_x}{1 - \alpha_A - \alpha_{BA} - \alpha_x})] \\ & + \alpha_x[-f(S - A - \{x\}) + f(S - B - \{x\}) \\ & - \lambda \log(1 + \frac{\alpha_{BA}}{1 - \alpha_A - \alpha_x - \alpha_{BA}})] \\ & + \alpha_A[-f(S - A - \{x\}) + f(S - B - \{x\}) + f(S - A) \\ & - f(S - B) + \lambda \log(\frac{1 - \alpha_A}{1 - \alpha_A - \alpha_x} \frac{1 - \alpha_A - \alpha_{BA} - \alpha_x}{1 - \alpha_A - \alpha_{BA}})] \\ & + 1[f(S - A - \{x\}) - f(S - B - \{x\}) - f(S - A) \\ & + f(S - B) + \lambda \log(\frac{1 - \alpha_A - \alpha_x}{1 - \alpha_A} \frac{1 - \alpha_A - \alpha_{BA}}{1 - \alpha_A - \alpha_{BA} - \alpha_x})] \\ & \geq 0 \end{aligned}$$

We assume $g(x+y) = f(x+y) - f(x)$, $q(x+y) = g(x+y) - g(x)$, $y \geq 0$. We can then rewrite the factorization as:

$$\begin{aligned} & \alpha_{BA}[-g(S - B) - \lambda \log(1 + \frac{\alpha_x}{1 - \alpha_A - \alpha_{BA} - \alpha_x})] \\ & + \alpha_x[-g(S - A - \{x\}) - \lambda \log(1 + \frac{\alpha_{BA}}{1 - \alpha_A - \alpha_x - \alpha_{BA}})] \\ & + \alpha_A[q(S - A) + \lambda \log(\frac{1 - \alpha_A}{1 - \alpha_A - \alpha_x} \frac{1 - \alpha_A - \alpha_{BA} - \alpha_x}{1 - \alpha_A - \alpha_{BA}})] \\ & + 1[-q(S - A) + \lambda \log(\frac{1 - \alpha_A - \alpha_x}{1 - \alpha_A} \frac{1 - \alpha_A - \alpha_{BA}}{1 - \alpha_A - \alpha_{BA} - \alpha_x})] \\ & \geq 0 \end{aligned}$$

Since the above inequality needs to hold true for all possible sets of $A \subseteq B$, $x \notin B$, we aim to determine the maximal amount enforced by the above set of inequalities. The first factorization results in two inequalities: (1) $g(C) \leq 0$ for all sets C and (2) $|g(C)| \geq \lambda \log(1 + \frac{\alpha_x}{1 - \alpha_B - \alpha_x})$. To find the maximum of such a limit, we need to impose the one item with highest probability to $\{x\}$ and assume that α_B is maximal while still less than $1 - \alpha_x$. This limit is imposed so that the denominator is not equal to 0.

Once again, it is important to note that by doing so, we are removing the possibility of $B = \{x\}^C$ as a case. In other words, we are assuming that $\alpha_B + \alpha_x = 1$ is not a scenario we need to discuss. We will now demonstrate why such an assumption is accurate.

When $\alpha_B + \alpha_x = 1$, we can rewrite inequality (8) in the following format:

$$\begin{aligned}
& (1 - \alpha_A - \alpha_x)f(S - A - \{x\}) + \\
& \lambda(1 - \alpha_A - \alpha_x) \log(1 - \alpha_A - \alpha_x) - (1 - \alpha_A)f(S - A) \\
& - \lambda(1 - \alpha_A) \log(1 - \alpha_A) \geq - (1 - \alpha_A - \alpha_{BA})f(S - B) \\
& - \lambda(1 - \alpha_A - \alpha_{BA}) \log(1 - \alpha_A - \alpha_{BA}) \rightarrow \\
& \alpha_{BA}\{-f(S - B) - \lambda \log(1 - \alpha_A - \alpha_{BA})\} + \\
& \alpha_x\{-f(S - A - \{x\}) - \lambda \log(1 - \alpha_A - \alpha_{BA})\} + \\
& \alpha_A\{-f(S - A - \{x\}) + f(S - A) - f(S - B) - \\
& \lambda \log(1 - \alpha_A - \alpha_x) + \lambda \log(1 - \alpha_A) \\
& - \lambda \log(1 - \alpha_A - \alpha_{BA})\} + 1\{f(S - A - \{x\}) - f(S - A) + \\
& f(S - B) + \lambda \log(1 - \alpha_A - \alpha_x) - \lambda \log(1 - \alpha_A) \\
& + \lambda \log(1 - \alpha_A - \alpha_{BA})\} \geq 0
\end{aligned}$$

By adding and removing three factors of $f(S - B - \{x\})$, $\lambda \log(1 - \alpha_A - \alpha_x)$ and $\lambda \log(1 - \alpha_A - \alpha_{BA})$ to each of the factorizations above and using the definitions of g and q functions as previously, we can simplify the above inequality in the following format:

$$\begin{aligned}
& \alpha_{BA}\{-g(S - B) + \lambda \log(1 - \alpha_A - \alpha_x)\} + \\
& \alpha_x\{-g(S - A - \{x\}) + \lambda \log(1 - \alpha_A - \alpha_{BA})\} + \\
& \alpha_A\{q(S - A) + \lambda \log(1 - \alpha_A)\} + \\
& + 1\{-q(S - A)\} - \lambda \log(1 - \alpha_A)\} + \\
& \{\lambda \log(1 - \alpha_A - \alpha_x) + \lambda \log(1 - \alpha_A - \alpha_{BA})\} \\
& \{-\alpha_{BA} - \alpha_x - \alpha_A + 1\} \geq 0
\end{aligned}$$

Using the fact that $\alpha_{BA} + \alpha_x + \alpha_A = 1$ and merging the 3rd and 4th factorizations together, we will have:

$$\begin{aligned}
& \alpha_{BA}\{-g(S - B) + \lambda \log(1 - \alpha_A - \alpha_x)\} + \\
& \alpha_x\{-g(S - A - \{x\}) + \lambda \log(1 - \alpha_A - \alpha_{BA})\} + \\
& (1 - \alpha_A)\{-q(S - A) - \lambda \log(1 - \alpha_A)\} \geq 0
\end{aligned}$$

which would hold true as long as (1) $g(C) \leq 0$ for all sets C and (2) $|g(C)| \geq \lambda \max(\log \frac{1}{\alpha_x})$ and (3) $q(C) \leq 0$ for all sets C . It turns out that the third condition will be required in all other instances of the set B as well meaning this specific case does not impose any severe condition changes. However, it does demand a different lower bound for the absolute value of the first order difference of function f over sets.

The second factorization results in two inequalities: (1) $g(C) \leq 0$ for all sets C and (2) $|g(C)| \geq \lambda \max(\log(1 + \frac{\alpha_{BA}}{1 - \alpha_B - \alpha_x}))$. To find the maximum of such a limit, we once again need to impose the one item with highest probability to $\{x\}$ and assume that $\alpha_A = 0$ and α_B is maximal while still less than $1 - \alpha_x$. This limit is once again imposed so that the denominator is not equal to 0.

The third and forth factorizations could be simplified. We could write them as

$$\begin{aligned}
& q(S - A)[-1 + \alpha_A] + \\
& \lambda \alpha_A \log\left(\frac{1 - \alpha_A}{1 - \alpha_A - \alpha_x} \frac{1 - \alpha_A - \alpha_{BA} - \alpha_x}{1 - \alpha_A - \alpha_{BA}}\right) \\
& + \lambda \log\left(\frac{1 - \alpha_A}{1 - \alpha_A - \alpha_{BA}} \frac{1 - \alpha_A - \alpha_x}{1 - \alpha_A - \alpha_{BA} - \alpha_x}\right) \geq 0
\end{aligned}$$

Depending on whether the sum of logarithmic arguments on the LHS is positive or negative, we can write two inequalities:

$$\begin{aligned}
(1) \quad & \lambda \alpha_A \log\left(\frac{1 - \alpha_A}{1 - \alpha_A - \alpha_x} \frac{1 - \alpha_A - \alpha_{BA} - \alpha_x}{1 - \alpha_A - \alpha_{BA}}\right) \\
& + \lambda \log\left(\frac{1 - \alpha_A}{1 - \alpha_A - \alpha_{BA}} \frac{1 - \alpha_A - \alpha_x}{1 - \alpha_A - \alpha_{BA} - \alpha_x}\right) \geq 0 \rightarrow \\
& q(S - A)[-1 + \alpha_A] + \\
& \lambda \alpha_A \log\left(\frac{1 - \alpha_A}{1 - \alpha_A - \alpha_x} \frac{1 - \alpha_A - \alpha_{BA} - \alpha_x}{1 - \alpha_A - \alpha_{BA}}\right) \\
& + \lambda \log\left(\frac{1 - \alpha_A}{1 - \alpha_A - \alpha_{BA}} \frac{1 - \alpha_A - \alpha_x}{1 - \alpha_A - \alpha_{BA} - \alpha_x}\right) \geq \\
& q(S - A)[-1 + \alpha_A] + \lambda \alpha_A \log\left(\frac{1 - \alpha_A}{1 - \alpha_A - \alpha_{BA}}\right) \stackrel{?}{\geq} 0
\end{aligned}$$

where we have simply summed the two logarithmic arguments. It then turns out that the logarithmic argument is always positive seeing as how the nominator of the fraction inside it is greater than the denominator. It thus follows that we merely need to impose that $q(C) \leq 0$ for all sets C to help above inequality hold true. A second scenario dictates when:

$$\begin{aligned}
(1) \quad & \lambda \alpha_A \log\left(\frac{1 - \alpha_A}{1 - \alpha_A - \alpha_x} \frac{1 - \alpha_A - \alpha_{BA} - \alpha_x}{1 - \alpha_A - \alpha_{BA}}\right) \\
& + \lambda \log\left(\frac{1 - \alpha_A}{1 - \alpha_A - \alpha_{BA}} \frac{1 - \alpha_A - \alpha_x}{1 - \alpha_A - \alpha_{BA} - \alpha_x}\right) \leq 0 \rightarrow \\
& q(S - A)[-1 + \alpha_A] + \\
& \lambda \alpha_A \log\left(\frac{1 - \alpha_A}{1 - \alpha_A - \alpha_x} \frac{1 - \alpha_A - \alpha_{BA} - \alpha_x}{1 - \alpha_A - \alpha_{BA}}\right) \\
& + \lambda \log\left(\frac{1 - \alpha_A}{1 - \alpha_A - \alpha_{BA}} \frac{1 - \alpha_A - \alpha_x}{1 - \alpha_A - \alpha_{BA} - \alpha_x}\right) \geq \\
& q(S - A)[-1 + \alpha_A] + \lambda \log\left(\frac{1 - \alpha_A}{1 - \alpha_A - \alpha_{BA}}\right) \stackrel{?}{\geq} 0
\end{aligned}$$

where we have once again simply summed the two logarithmic arguments. It again turns out that the logarithmic argument is always positive seeing as how the nominator of the fraction inside it is greater than the denominator. It once more thus follows that we merely need to impose that $q(C) \leq 0$ for all sets C to help above inequality hold true. Thus 3 sufficient conditions for submodularity of $F_2(S_j), j = 1, \dots, N$ are developed: (1) $g(C) \leq 0$ (2) $q(C) \leq 0$ (3) $|g(C)| \geq \lambda \log[1 + \frac{\pi_1}{\pi_M}, 1 + \frac{1 - \pi_M - \pi_1}{\pi_M}]$.

Thus, the total set of sufficient conditions for submodularity of a general M -to- N binning problem with utility functions as described previously is developed as follows:

- (1) $g(C) \leq 0$
- (2) $q(C) \leq 0$

(3) $|g(C)| \geq \lambda \log(1 + \max(\frac{\pi_1}{\pi_M}, \frac{1-\pi_M-\pi_1}{\pi_M}, \frac{1-\pi_M}{\pi_M})) = \lambda \log(\frac{1}{\pi_M})$
for all sets C .

□

C. Lemma III.3

Proof. For the specific case $N = 2$:

$$F(A) = \alpha_A f(A) + (1 - \alpha_A) f(S - A) + \alpha_A \log(\alpha_A) + (1 - \alpha_A) \log(1 - \alpha_A)$$

In such a case, diminishing return property requires that

$$\begin{aligned} &(\alpha_A + \alpha_x) f(A + \{x\}) + (1 - \alpha_A - \alpha_x) f(S - A - \{x\}) \\ &+ (\alpha_A + \alpha_x) \log(\alpha_A + \alpha_x) \\ &+ (1 - \alpha_A - \alpha_x) \log(1 - \alpha_A - \alpha_x) - (\alpha_A) f(A) \\ &- (1 - \alpha_A) f(S - A) - (\alpha_A) \log(\alpha_A) \\ &- (1 - \alpha_A) \log(1 - \alpha_A) \geq \\ &(\alpha_A + \alpha_{BA} + \alpha_x) f(B + \{x\}) \\ &+ (1 - \alpha_A - \alpha_{BA} - \alpha_x) f(S - B - \{x\}) \\ &+ (\alpha_A + \alpha_{BA} + \alpha_x) \log(\alpha_A + \alpha_{BA} + \alpha_x) \\ &+ (1 - \alpha_A - \alpha_{BA} - \alpha_x) \log(1 - \alpha_{BA} - \alpha_A - \alpha_x) \\ &- (\alpha_B) f(B) - (1 - \alpha_A - \alpha_{BA}) f(S - B) \\ &- (\alpha_A + \alpha_{BA}) \log(\alpha_A + \alpha_{BA}) \\ &- (1 - \alpha_A - \alpha_{BA}) \log(1 - \alpha_A - \alpha_{BA}) \end{aligned}$$

We then factorize this inequality in the following manner:

$$\begin{aligned} &\alpha_{BA} \{-f(B + \{x\}) + f(B) + f(S - B - \{x\}) - f(S - B) \\ &+ \lambda \log(\frac{\alpha_A + \alpha_{BA}}{\alpha_A + \alpha_{BA} + \alpha_x} \frac{1 - \alpha_A - \alpha_{BA} - \alpha_x}{1 - \alpha_A - \alpha_{BA}})\} + \\ &\alpha_x \{f(A + \{x\}) - f(B + \{x\}) - f(S - A - \{x\}) \\ &+ f(S - B - \{x\}) + \\ &\lambda \log(\frac{\alpha_A + \alpha_x}{\alpha_A + \alpha_x + \alpha_{BA}} \frac{1 - \alpha_A - \alpha_x - \alpha_{BA}}{1 - \alpha_A - \alpha_x})\} \\ &+ \alpha_A \{f(A + \{x\}) - f(B + \{x\}) - f(S - A - \{x\}) \\ &+ f(S - B - \{x\}) - f(A) + f(B) + f(S - A) - f(S - B) \\ &+ \log(\frac{\alpha_A + \alpha_x}{\alpha_A + \alpha_x + \alpha_{BA}} \frac{1 - \alpha_A - \alpha_x - \alpha_{BA}}{1 - \alpha_A - \alpha_x} \frac{\alpha_A + \alpha_{BA}}{\alpha_A} \\ &\frac{1 - \alpha_A}{1 - \alpha_A - \alpha_{BA}})\} \\ &+ \{f(S - A - \{x\}) - f(S - B - \{x\}) - f(S - A) \\ &+ f(S - B) + \lambda \log(\frac{1 - \alpha_A - \alpha_x}{1 - \alpha_A - \alpha_x - \alpha_{BA}} \frac{1 - \alpha_A - \alpha_{BA}}{1 - \alpha_A})\} \\ &\geq 0 \end{aligned}$$

We now introduce two new set functions $g(C + D) = f(C + D) - f(C)$ and $q(C + D) = g(C + D) - g(C)$ where C, D represent two arbitrary sets belonging to the universal

set. We then use these new definitions to simplify the above inequality:

$$\begin{aligned} &\alpha_{BA} \{-g(B + \{x\}) - g(S - B) \\ &+ \lambda \log(\frac{\alpha_A + \alpha_{BA}}{\alpha_A + \alpha_{BA} + \alpha_x} \frac{1 - \alpha_A - \alpha_{BA} - \alpha_x}{1 - \alpha_A - \alpha_{BA}})\} + \\ &\alpha_x \{-g(B + \{x\}) - g(S - A - \{x\}) + \\ &\lambda \log(\frac{\alpha_A + \alpha_x}{\alpha_A + \alpha_x + \alpha_{BA}} \frac{1 - \alpha_A - \alpha_x - \alpha_{BA}}{1 - \alpha_A - \alpha_x})\} \\ &+ \alpha_A \{-q(B + \{x\}) + q(S - A) + \\ &\log(\frac{\alpha_A + \alpha_x}{\alpha_A + \alpha_x + \alpha_{BA}} \frac{1 - \alpha_A - \alpha_x - \alpha_{BA}}{1 - \alpha_A - \alpha_x} \frac{\alpha_A + \alpha_{BA}}{\alpha_A} \\ &\frac{1 - \alpha_A}{1 - \alpha_A - \alpha_{BA}})\} \\ &+ \{-q(S - A) + \lambda \log(\frac{1 - \alpha_A - \alpha_x}{1 - \alpha_A - \alpha_x - \alpha_{BA}} \frac{1 - \alpha_A - \alpha_{BA}}{1 - \alpha_A})\} \\ &\geq 0 \end{aligned}$$

We aim to find sufficient conditions so that the above inequality can hold true. The first factorization is the coefficients of $\alpha_{BA} \geq 0$. We thus need to ascertain that the coefficient is also positive. To do so, we impose that

$$-2g(C) + \lambda \log(\frac{\alpha_A + \alpha_{BA}}{\alpha_A + \alpha_{BA} + \alpha_x} \frac{1 - \alpha_A - \alpha_{BA} - \alpha_x}{1 - \alpha_A - \alpha_{BA}}) \geq 0$$

for all possible sets C . We thus need to impose two sufficient conditions:

$$(1) \quad g(C) \leq 0, \forall C$$

$$(2) \quad 2|g(C)| \geq$$

$$\max(|\lambda \log(\frac{\alpha_A + \alpha_{BA}}{\alpha_A + \alpha_{BA} + \alpha_x} \frac{1 - \alpha_A - \alpha_{BA} - \alpha_x}{1 - \alpha_A - \alpha_{BA}})|)$$

The second factorization is the coefficients of $\alpha_x \geq 0$. We once again need to ascertain that the coefficient is also positive. To do so, we impose that

$$(1) \quad g(C) \leq 0, \forall C$$

$$(2) \quad 2|g(C)| \geq$$

$$\max(|\lambda \log(\frac{\alpha_A + \alpha_x}{\alpha_A + \alpha_x + \alpha_{BA}} \frac{1 - \alpha_A - \alpha_x - \alpha_{BA}}{1 - \alpha_A - \alpha_x})|)$$

We could see that the RHS of the secondary condition could be written as:

$$\lambda \max(\log((1 + \frac{\alpha_{BA}}{\alpha_A + \alpha_x})(1 + \frac{\alpha_{BA}}{1 - \alpha_A - \alpha_x}))) = K \quad (10)$$

It is clear to see that the two fractions could not be maximal at the same time seeing as how (1) the maximal occurs when denominator of each is close to zero and (2) their denominators have opposing behavior. It thus turns out that the RHS as indicated in Eq. (10) is limited by:

$$K < \lambda \log(\max(1 + \frac{\alpha_{BA}}{\alpha_A + \alpha_x}) \max(1 + \frac{\alpha_{BA}}{1 - \alpha_A - \alpha_x}))$$

The first fraction is maximized when $\alpha_A = 0$, $\alpha_x = \pi_M$ and $\alpha_{BA} = 1 - \pi_M$ while the second fraction can never be as high. It thus follows that:

$$K < \lambda \log \left(\frac{1}{\pi_M} \right)^2$$

□